Long-term Archive Challenges: Enhancing Data Discovery via Multilevel Metadata Aggregations At Scale

Graham Parton, Ag Stephens, Richard Smith, Joe Singleton
Centre for Environmental Data Analysis (CEDA), STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot, United Kingdom (graham.parton@stfc.ac.uk)

Photography encouraged

Introduction

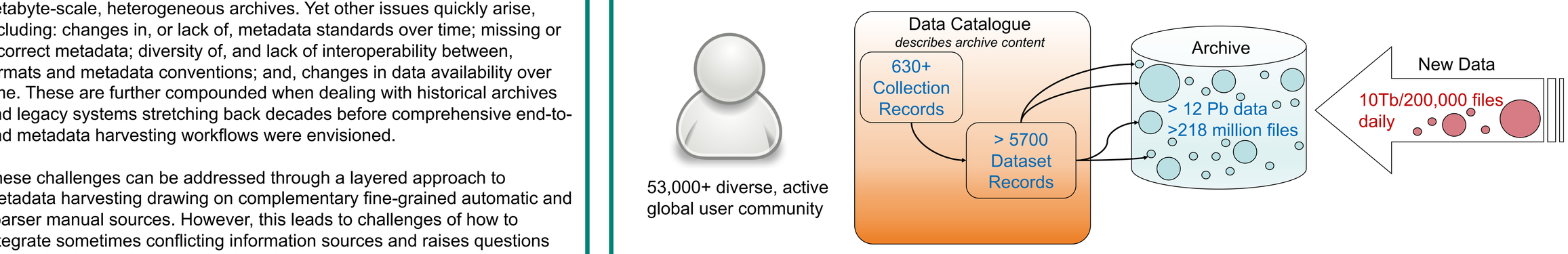
Data archives require accurate, content-rich data catalogues that are fit-for-purpose to support meaningful data discovery. However, sourcing suitable high-quality metadata to populate catalogues at scale can be problematic when manual workflows are no longer able to cope. One solution is automated metadata harvesting directly from data files, drawing on technical solutions to Big Data challenges faced by rapidly evolving, petabyte-scale, heterogeneous archives. Yet other issues quickly arise, including: changes in, or lack of, metadata standards over time; missing or incorrect metadata; diversity of, and lack of interoperability between, formats and metadata conventions; and, changes in data availability over time. These are further compounded when dealing with historical archives and legacy systems stretching back decades before comprehensive end-to-end metadata harvesting workflows were envisioned.

These challenges can be addressed through a layered approach to metadata harvesting drawing on complementary fine-grained automatic and coarser manual sources. However, this leads to challenges of how to integrate sometimes conflicting information sources and raises questions about what is 'truth' and where should it be conveyed.

CEDA Archive: The Big Data Challenges

The Centre for Environmental Data Analysis (CEDA) archive faces all 4 Big Data 'V' challenges:

- Volume – both in archive volumes and data catalogue records
- Velocity – in speeds of data arrival and required cataloguing changes
- Veracity – handling wide range of metadata and file variance and ensuring data preservation
- Variety – archiving the Atmospheric, Climate Change, Earth Observation data, 1000s of formats



At this scale producing and maintaining quality data catalogue content manually is no longer possible.

Big Data (partial) Solutions

The CEDA Archive is held within the UK's unique high-performance data analysis JASMIN platform. Utilising JASMIN's parallel processing power archive-wide scanning and metadata harvesting is possible on short, useful timescales at the individual file level. The content are then stored within an Elasticsearch index – a NoSQL content store – the CEDA File Based Index (FBI). This is updated with fresh content as files are ingested into the archive, providing a fine-grained, timely and highly scalable metadata harvesting solution to the Big Data challenges.

However, this is unable to manage all cases, such as:

- Offline, removed or external content to be catalogued
- Files not yielding metadata content. E.g. format is not scannable, or file lacks required metadata
- Files yielding incorrect metadata

CEDA File Based Index

The CEDA FBI holds a range of information scraped from each file within the CEDA archive. With the range of file formats in the CEDA archive it has not been possible to produce parsers for all file types to extract internal metadata, but basic information is available for all files. The full range of possible details stored per file include:

- Basic file information:
  - File size
  - Location
  - File extension
  - format information where possible
  - MD5 checksums
- Internal metadata – for 9 of the most common file formats parameter information is scanned including: standard names, long names, variable ids, units

CEDA are also in the process of developing tools to also extract:

- Geospatial bounding box
- Temporal range

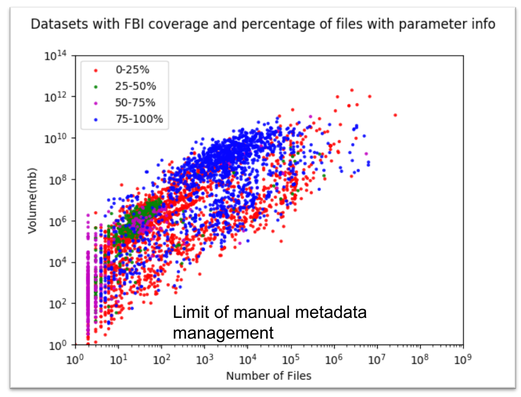


Fig 1. CEDA dataset level metadata aggregations from FBI

FBI limitations and Scalability

Whilst the FBI gives unprecedented metadata harvesting at scale and fast aggregations to dataset level, figure 1 indicates the limitations of this approach – whilst parameter information can be scraped from 48% of the archive it is not evenly distributed amongst datasets. Some datasets are metadata rich, whilst others remain sparsely documented. At the same time the distribution of datasets in terms of size and number of files presents a hard limit to manual approaches to metadata harvesting. Thus, there is a need to provide a complementary metadata source to cover the remaining datasets and where automated metadata harvest returns are low. Fortunately, CEDA has been curating such metadata for around 20 years, but a systematic service to store this information for complementary harvesting was required: the CEDA Manual Metadata Store.

CEDA Manual Metadata Store

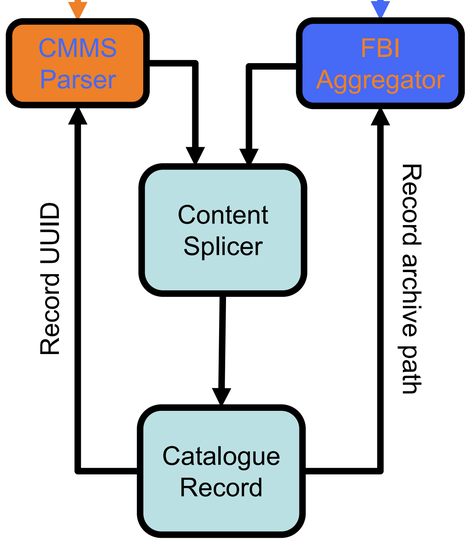
https://github.com/cedadev/cmms https://github.com/gap736uk/pyCMMS

The CMMS is designed as an easy to use and maintainable service for data scientists to manually curate metadata at the dataset level to augment aggregated content from the FBI. To ensure ease of use it:

- Is a web based store in GitHub (ease of versioning and inbuilt editor)
- Follows a simple YAML syntax
- Is complemented by a content parsing Python library pyCMMS
- pyCMMS also provides checking tool for CMMS content validation.

A CMMS YAML entry may contain the following content:

- Splice rules – indicating how CMMS and FBI content should be joined
- Parameter listings
- Temporal range
- Geographic Bounding Box
- Total Volume – for external, offline or removed content
- Number of files – for external, offline or removed content
- Licence and access details – for external content



Present Status and Future Implications

With the CMMS presently holding > 1000 entries and the pyCMMS library now available recent work on the CEDA catalogue has integrated the full suite of CMMS and FBI aggregations and content splicing ready for deployment. Initial tests on parameter harvesting are promising, indicating nearly complete coverage for CEDA dataset records (fig 2).

Fig 2. CEDA dataset level metadata CMMS coverage compared against FBI parameter coverage

However, with the FBI content now being available for the community to directly poll and CEDA's development of OpenSearch tools to provide facetted search based on the FBI new issues begin to emerge. In particular, there are known cases of incorrect or missing file-derived metadata within the FBI, though the CMMS content may not permit discrimination by file-type within a given dataset to be able to inject directly into the FBI.

Even if the CMMS content were further refined to apply on a per-file basis, questions then arise as to where to retrospectively apply the change (just in the FBI or adjust the file content themselves?) and, more fundamentally, *should* such changes be made or should we simply make users aware of known problems?

Metadata Splicing

Key to handling content from two metadata sources is controlling how the information is 'spliced' together. Within CEDA's workflows content from the FBI remains agnostic to the existence of CMMS entries and so rules exist within CMMS entries to control the content splicing as handled within the CEDA data catalogue's metadata harvesting tools. The permitted options are:

	CMMS Rule	Splicing Function
F else C	Default	Take FBI content if it exists, else default to CMMS entry
C	CMMS Only	Use CMMS entry alone, regardless of FBI returns.
F C	Append	FBI returns are augmented with new content only from CMMS
F C	Replace and append	FBI returns are used and augmented with CMMS returns. Where overlap exists CMMS entries are taken as 'truth'.
F C	Replace	FBI returns are used but replaced with 'truth' from CMMS where overlap exists.

Whilst other cases could also exist in theory, real-world work-flows limit the required CMMS splicing options to those above. Also, not all rules are applicable to all content types where the use-cases only require the 'Default' and 'CMMS Only' rules would apply, e.g. for total volume and file numbers.

Acknowledgements

The authors would like to express their gratitude to their colleagues in the Centre for Environmental Data Analysis in assisting in the ongoing development of the CEDA data catalogue, archive content and curation of additional metadata resources over the years. Additionally, it is only thanks to the work of data providers to adopt best-working practices and seek to archive data with rich, harvestable metadata within suitable formats that permits metadata harvesting at scale.

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL